

Distributed Information Filtering System for Digital Libraries

**Indiana University Purdue University Indianapolis
Indiana University Bloomington**


**Mathew J. Palakal
Snehasis Mukhopdhyay
Javed Mostafa
Rajeev Raje**

<http://sifter.indiana.edu>



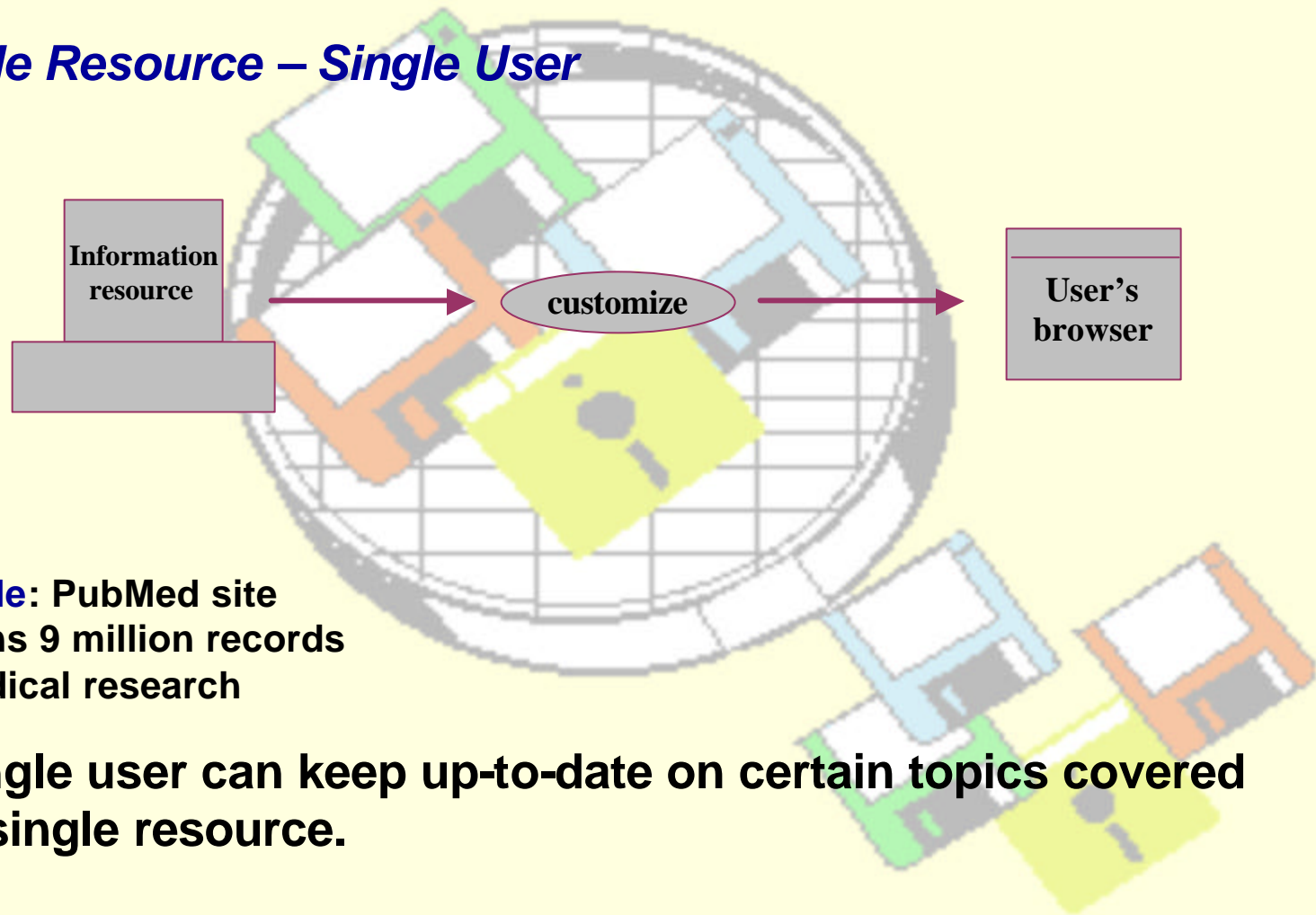
Motivation

"Filters are programs that select and prioritize information according to the instructions, needs, or customs of a given user. In a world increasingly filled with a flood of information and with users strained for available time, filters will assume an important role in the acquisition of information" [NSF Report on Research Priorities in Networking, 1994]



Motivation

Single Resource – Single User



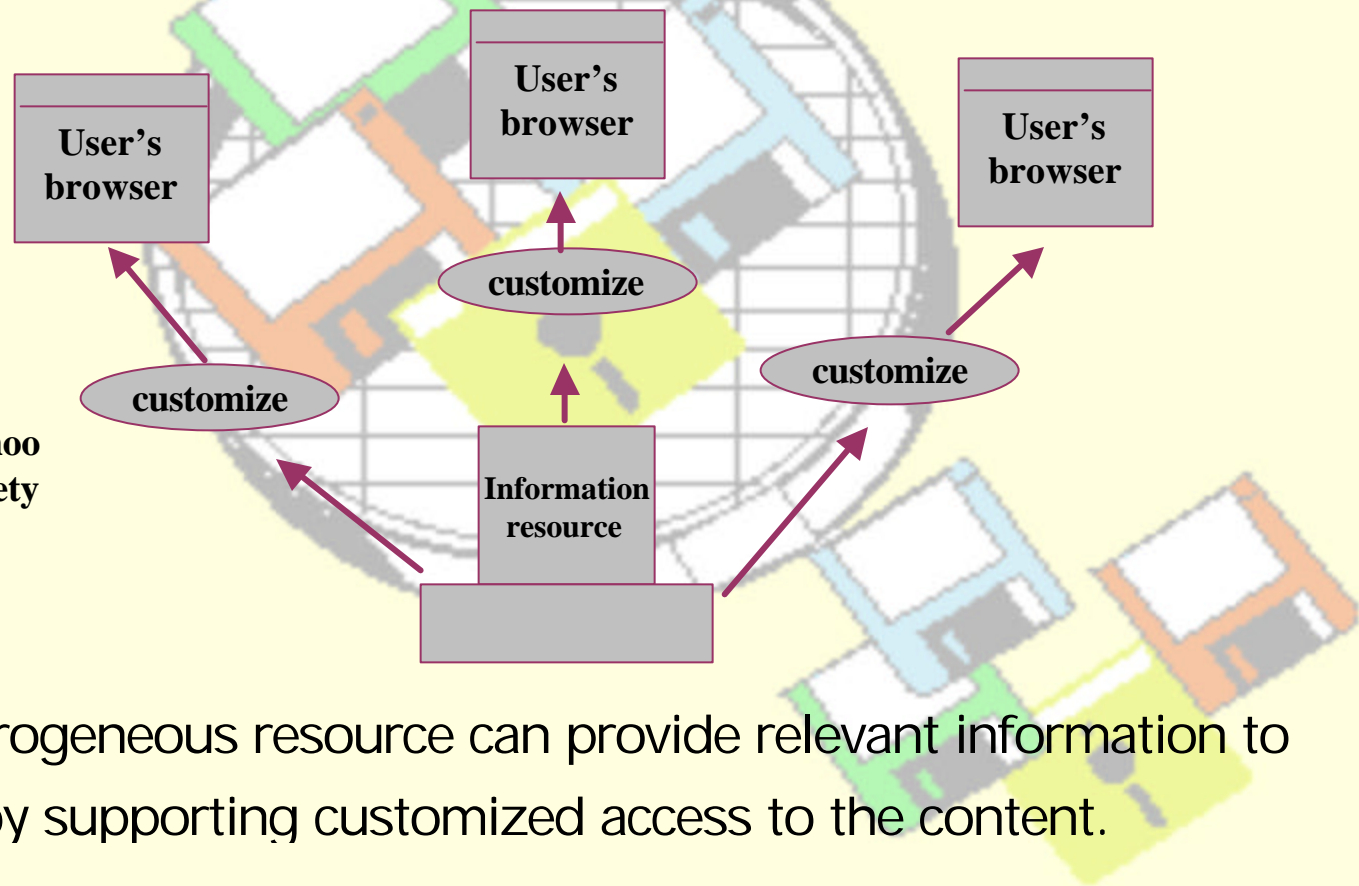
Example: PubMed site
contains 9 million records
on medical research

**A single user can keep up-to-date on certain topics covered
in a single resource.**

Motivation

Single Resource – Multiple Users

Example: A large portal such as Yahoo covers a wide variety of topics.

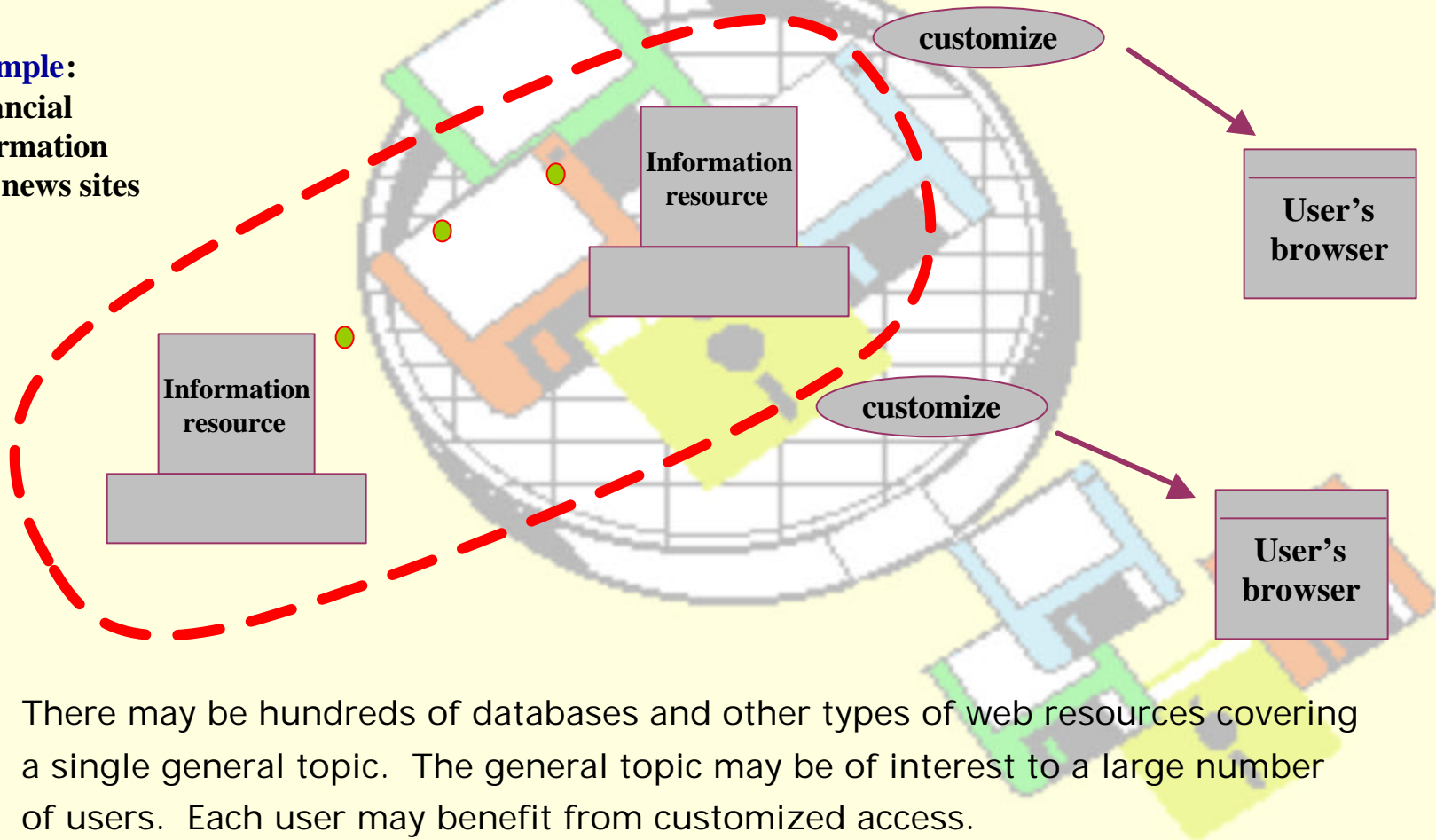


A large heterogeneous resource can provide relevant information to individuals by supporting customized access to the content.

Motivation

Multiple Resources – Multiple Users

Example:
Financial
information
and news sites



There may be hundreds of databases and other types of web resources covering a single general topic. The general topic may be of interest to a large number of users. Each user may benefit from customized access.

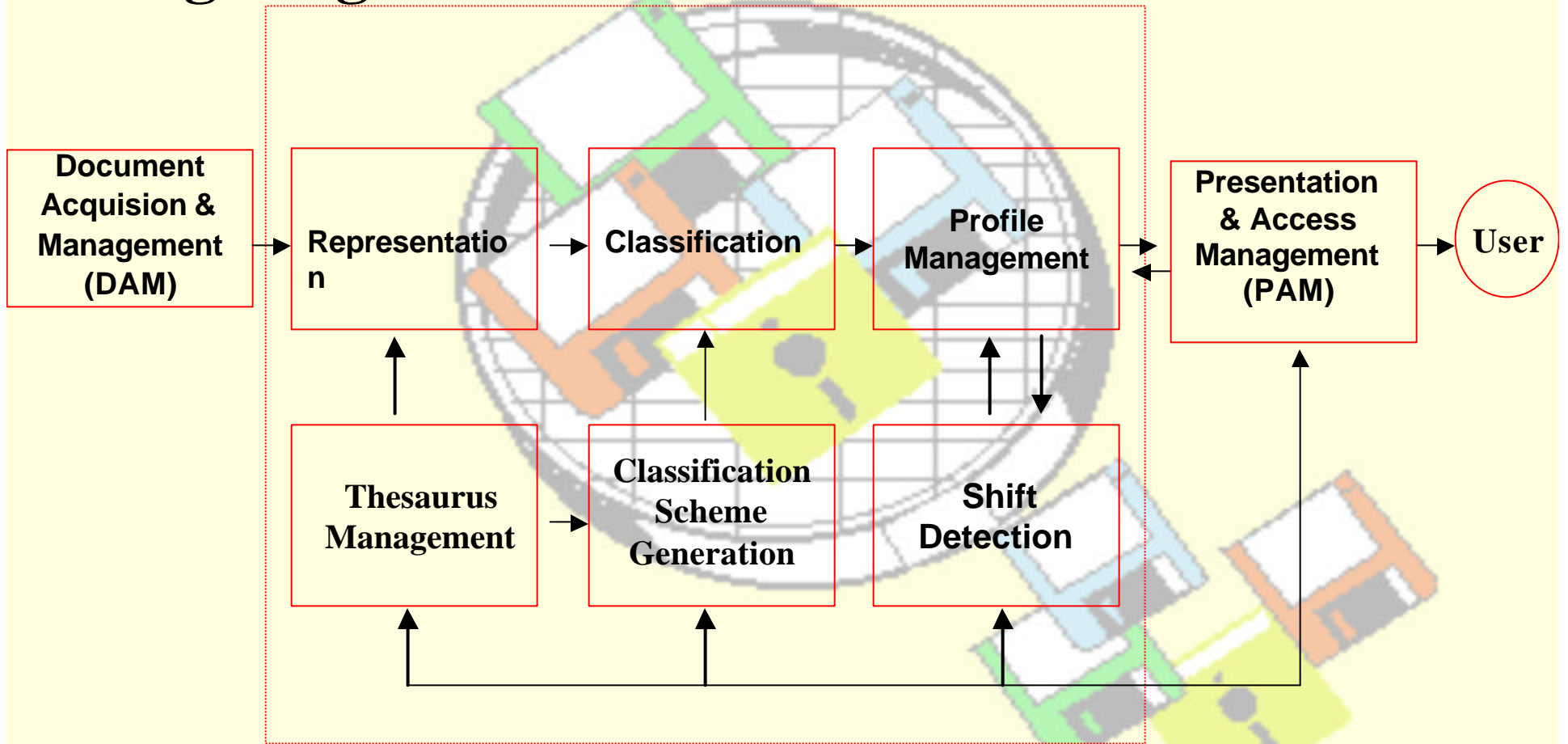
Objectives

**“Smart•Information•Filtering•Technology for
Electronic•Resources -SIFTER”**

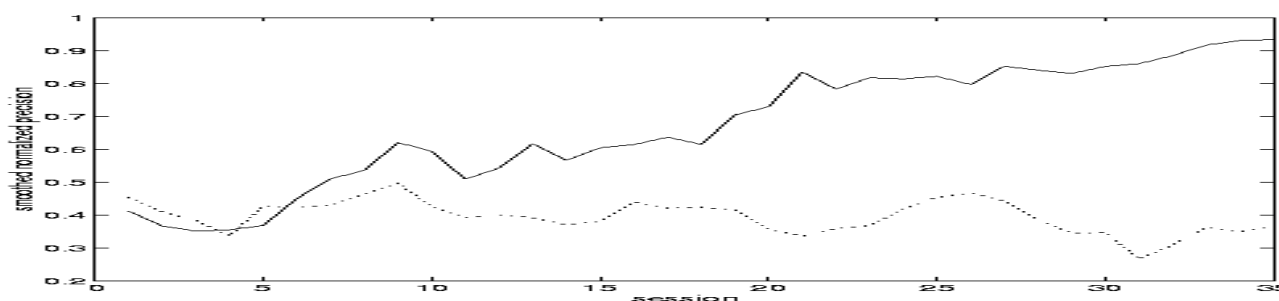
- **Reduce information overload**
- **Personalized access to information**
- **Handle varying user interest**
- **Provide reactive and proactive capabilities**
- **Achieve domain independency/efficiency through agent collaboration**

Prior Work

Single Agent Filter

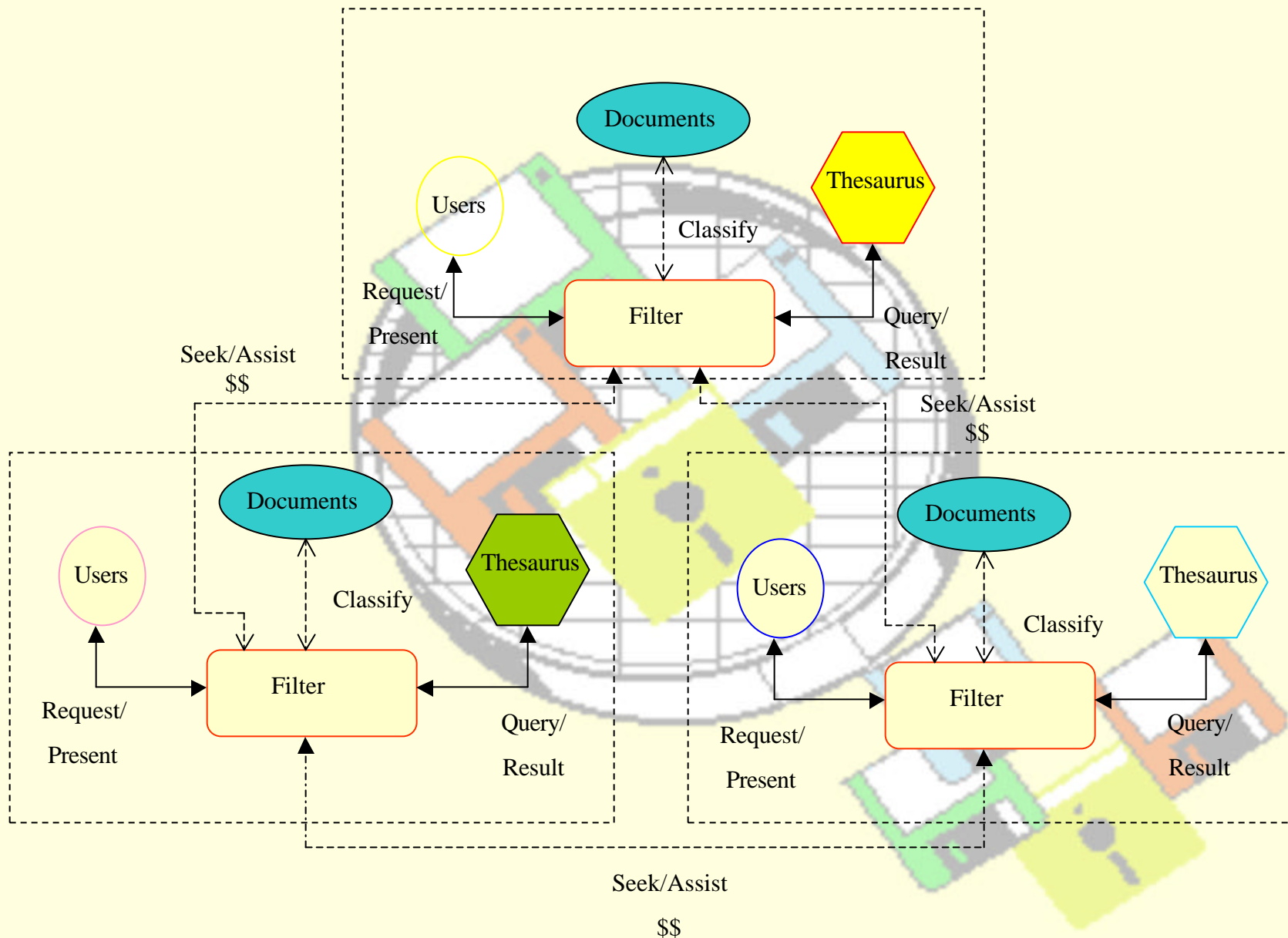


Prior Work: Single Agent Filter



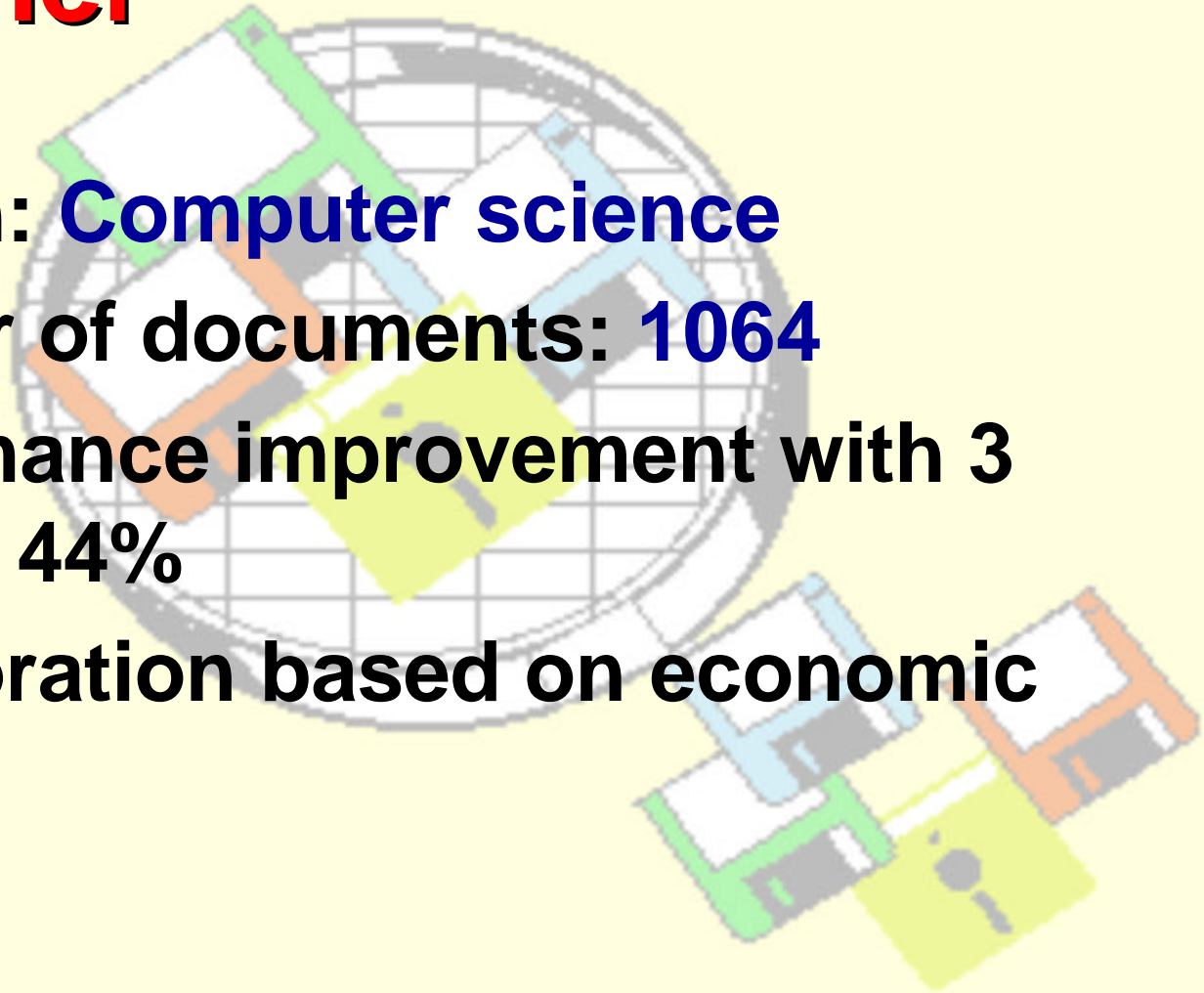
- Domain: **Computer Science**
- Number of documents: **6000**
- Experiments with real & simulated users

Prior work: Distributed Classifier



Prior work: Distributed Classifier

- Domain: **Computer science**
- Number of documents: **1064**
- Performance improvement with 3 agents: **44%**
- Collaboration based on economic model



Current Work: Distributed Filter

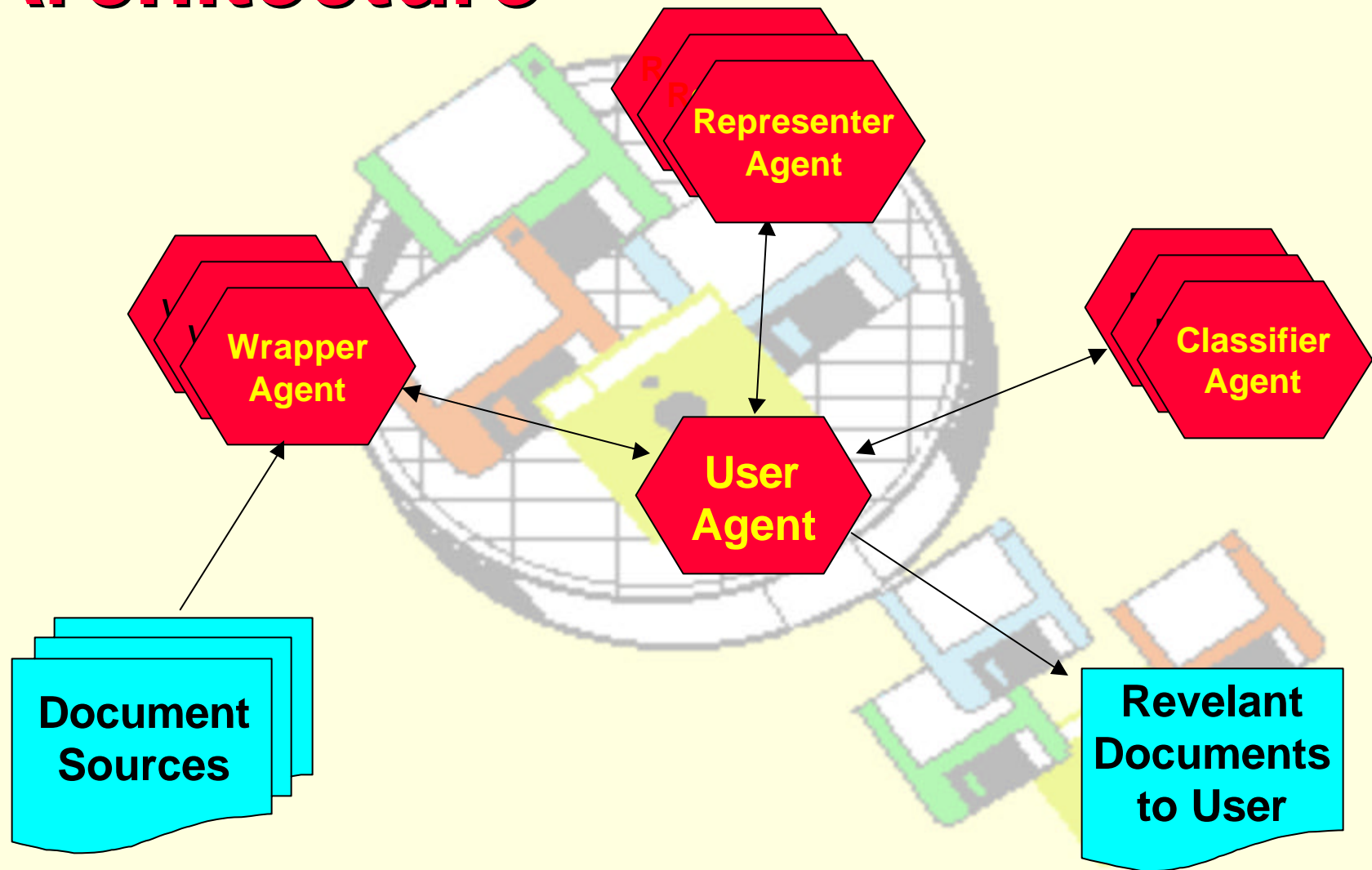
- Each filter consists of Thesaurus, Representation, Classification, Profiling, Collaboration
- All filters are identical (except thesaurus)
- Several collaboration schemes
 - single-opinion, multiple-opinion, combination
 - *economic(money-centric, user-centric, bidding)*

- Results

	Filter 1	Filter 2	Filter 3
Single Filter	326	0	0
Multi Filter	326	50	97

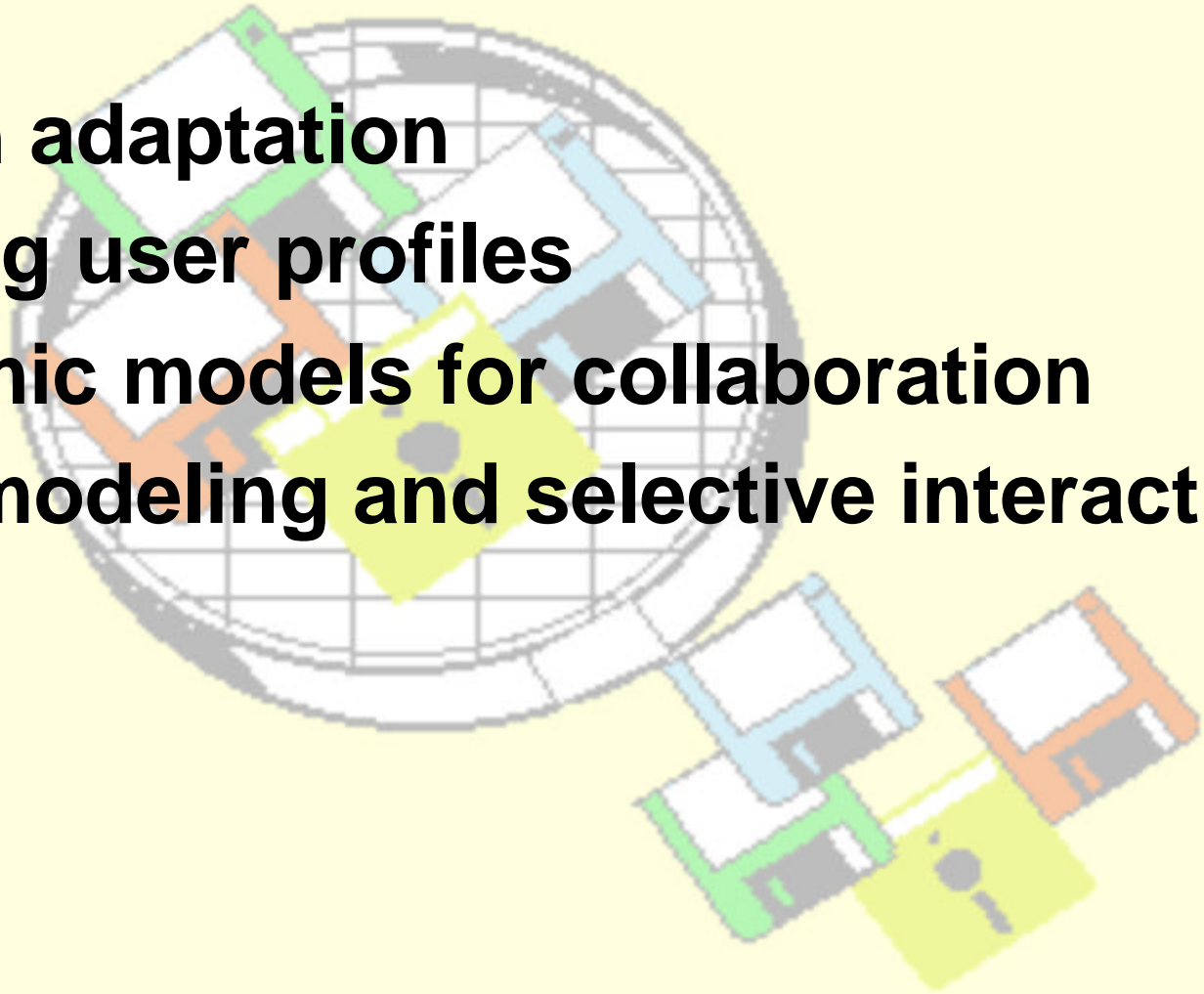
$N_{\text{Class}}_{\text{single}} = 12$ $N_{\text{Class}}_{\text{collaboration}} = 35$

Current Work: Multi-agent Architecture



Future Work

- **Domain adaptation**
- **Learning user profiles**
- **Economic models for collaboration**
- **Agent modeling and selective interaction**



Future Work: Domain Adaptation

- **Basic filtering model is general**
- **Domain dependencies:**
 - Thesaurus or vocabulary
 - Document base for representation & clustering
 - Classification space
- **Approaches to domain adaptation**
 - Multi-agent collaboration
 - Automated term discovery
 - Hierarchical thesaurus organization & classification
 - Centralized classification server vs. decentralized classification

Future Work: Learning User Profiles

- **Multi-level (fast but limited accuracy) vs. direct (slow but accurate)**
- **Approaches for direct learning:**
 - Neural nets (feedback on-line, train off-line)
 - Genetic learning (on-line)
 - Decision trees (discretization of features, off-line learning)
- **Hybrid approach: starting with multi-level and gradual transition to direct models**

Future Work: Economic Framework for Collaboration

- **Bidding function based on**
 - quality of service & financial state
 - probability of seeking remote collaboration
 - user-centric vs. money-centric
- **Four stage contracting (request/bid/select/execute) vs. more stages of negotiation**
- **Auction-based contracting**
- **Sub-contracting**

Future Work: Agent Modeling & Selective Interaction

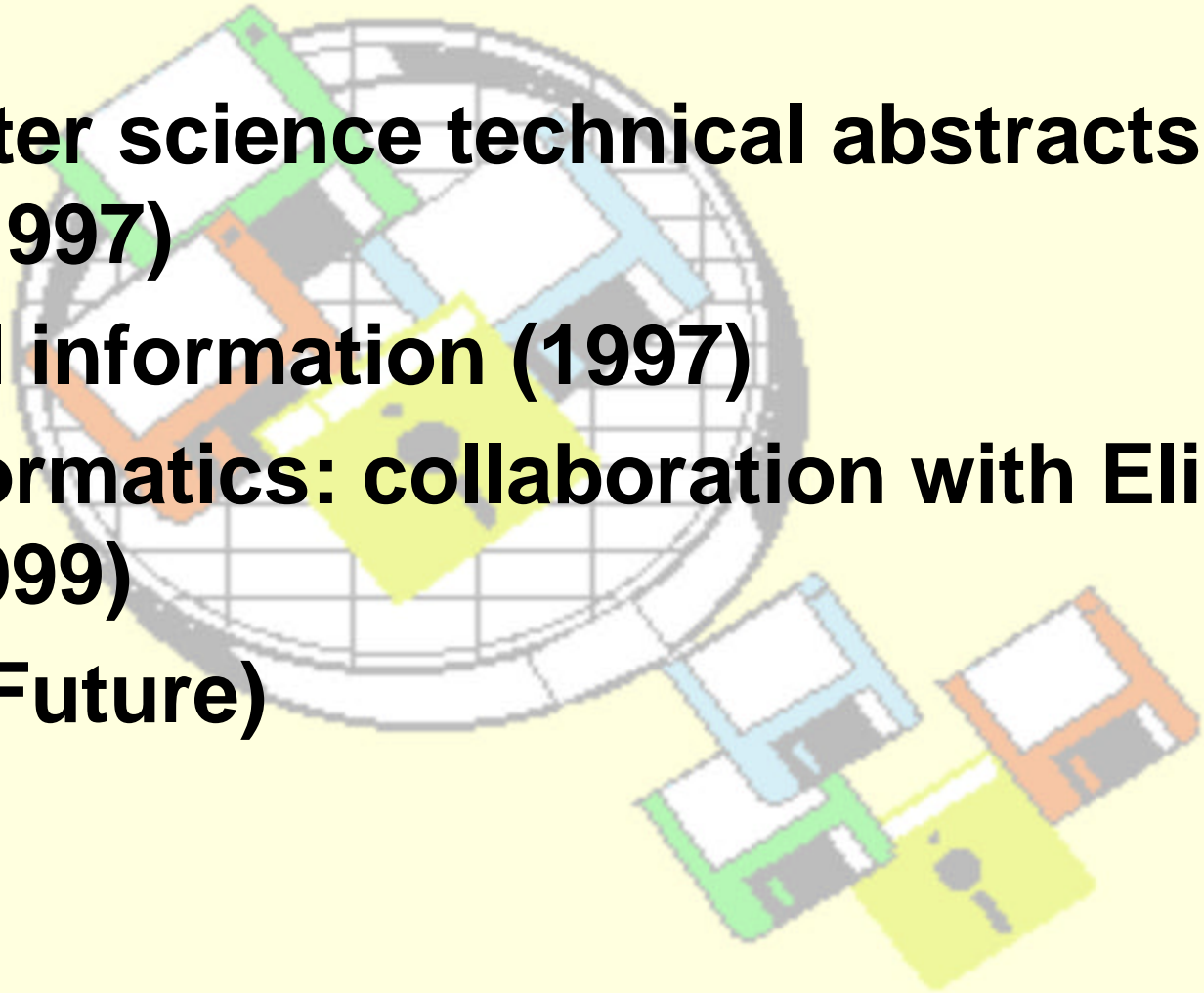
- **Large scale agent systems prohibit exhaustive interaction**
- **Models of other agents support selective interaction**
- **Acquaintance lists and graph**
- **Effect of selective interaction on system efficiency**

SIFTER Implementation

- **Single-agent filter (SIFTER)**
 - implemented in: **C & Tcl/Tk * Java**
- **Java demo accessible through Web browsers**
- **Multi-agent filter (D-SIFTER)** implemented in
 - **Java-RMI**
 - **CORBA (future implementation)**

Applications

- **Computer science technical abstracts (1996, 1997)**
- **Medical information (1997)**
- **Bio-informatics: collaboration with Eli Lilly (1999)**
- **TREC (Future)**



Salient Features of SIFTER

- High degree of performance with minimal user intervention
- Multi-level learning
- Adaptable to changes in document stream
- Adaptable to changes in user's interest
- Distributed agents collaborating using market-based economic models
- Implementation suited for heterogeneous, networked environments